Using a Non-Commutative Bernstein Bound to Approximate Some Matrix Algorithms in the Spectral Norm

Malik Magdon-Ismail
CS Department, Rensselaer Polytechnic Institute,
Troy, NY 12180, USA.
magdon@cs.rpi.edu

January 18, 2013

Abstract

We focus on *row sampling* based approximations for matrix algorithms, in particular matrix multipication, sparse matrix reconstruction, and ℓ_2 regression. For $\mathrm{A} \in \mathbb{R}^{m\times d}$ (m points in $d \ll m$ dimensions), and appropriate row-sampling probabilities, which typically depend on the norms of the rows of the $m \times d$ left singular matrix of A (the *leverage scores*), we give row-sampling algorithms with linear (up to polylog factors) dependence on the stable rank of A. This result is achieved through the application of non-commutative Bernstein bounds.

Keywords: row-sampling; matrix multiplication; matrix reconstruction; estimating spectral norm; linear regression; randomized

1 Introduction

Matrix algorithms (eg. matrix multiplication, SVD, ℓ_2 regression) are of widespread use in many application areas: data mining (Azar *et al.*, 2001); recommendations systems (Drineas *et al.*, 2002); information retrieval (Berry *et al.*, 1995; Papadimitriou *et al.*, 2000); web search (Kleinberg, 1999; Achlioptas *et al.*, 2001); clustering (Drineas *et al.*, 2004; McSherry, 2001); mixture modeling (Kannan *et al.*, 2008; Achlioptas and McSherry, 2005); etc. Based on the importance of matrix algorithms, there has been considerable research energy expended on breaking the $O(md^2)$ bound required by exact SVD methods (Golub and van Loan, 1996).

Starting with a seminal result of Frieze *et al.* (1998), a large number of results using non-uniform sampling to speed up matrix computations have appeared (Achlioptas and McSherry, 2007; Deshpande *et al.*, 2006; Deshpande and Vempala, 2006; Drineas *et al.*, 2006a,b,c,d,e; Rudelson and Vershynin, 2007; Magen and Zouzias, 2010), some of which give relative error guarantees (Deshpande *et al.*, 2006; Deshpande and Vempala, 2006; Drineas *et al.*, 2006d,e; Magen and Zouzias, 2010).

Even more recently, Sarlos (2006) showed how random projections or "sketches" can be used to perform all these tasks efficiently, obtaining the first $o(md^2)$ algorithms when preserving the identity of the rows themselves are not important. In fact, we will find many of these techniques, together with those in Ailon and Chazelle (2006) essential to our algorithm for generating row samples ultimately leading to $o(md^2)$ algorithms based on row-sampling. From now on, we focus on row-sampling algorithms.

We start with the basic result of matrix multiplication. All other results more or less follow from here. In an independent recent work which is developed along the lines of using isoperimetric inequalities (Rudelson and Vershynin, 2007) to obtain matrix Chernoff bounds, Magen and Zouzias

(2010) show that by sampling nearly a linear number of rows, it is possible to obtain a relative error approximation to matrix multiplication. Specifically, let $\mathrm{A} \in \mathbb{R}^{m \times d_1}$ and $\mathrm{B} \in \mathbb{R}^{m \times d_2}$. Then, for $r = \Omega(\rho/\epsilon^2 \log(d_1 + d_2))$ (where ρ bounds the stable (or "soft") rank of A and B – see later), there is a probability distribution over $\mathcal{I} = \{1, \ldots, m\}$ such that by sampling r rows i.i.d. from $\mathcal{I}$, one can construct sketches $\tilde{\mathrm{A}}$, $\tilde{\mathrm{B}}$ such that $\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{B}} \approx \mathrm{A}^{\mathrm{T}}\mathrm{B}$. Specifically, with constant probability,

$$\|\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{B}} - \mathrm{A}^{\mathrm{T}}\mathrm{B}\|_2 \leq \epsilon \|\mathrm{A}\|_2 \|\mathrm{B}\|_2.$$

The sampling distribution is relatively simple, relying only on the product of the norms of the rows in A and B. This result is applied to low rank matrix reconstruction and ℓ_2-regression where the required sampling distribution needs knowledge of the SVD of A and B.

Our basic result for matrix multiplication is very similar to this, and we arrive at it through a different path using a non-commutative Bernstein bound. Our sampling probabilities are different. In appication of our results to sparse matrix reconstruction and ℓ_2-regression, the rows of the left singular matrix make an appearance. In Magdon-Ismail (2010), it is shown how to approximate these probabilities in $o(md^2)$ time using random projections at the expense of a poly-logarithmic factor in running times. Further refinements lead to an even more efficient algorithm Drineas *et al.* (2010). As mentioned above, we must confess that one may perform our matrix tasks more efficiently using these same random projection methods (Sarlos, 2006), however the resulting algorithms are in terms of a small number of linear combinations of all the rows. In many applications, the actual rows of A have some physical meaning and so methods based on a small number of the actual rows are of interest.

We finally mention that Magen and Zouzias (2010) also give a dimension independent bound for matrix multiplication using some stronger tools. Namely, one can get the matrix multiplication approximation in the spectral norm using $r = \Omega(\rho/\epsilon^2 \log(\rho/\epsilon^2))$. In practice, it is not clear which bound is better, since there is now an additional factor of $1/\epsilon^2$ inside the logarithm.

1.1 Basic Notation

Before we can state the results in concrete form, we need some preliminary conventions. In general, $\epsilon \in (0, 1)$ will be an error tolerance parameter; $\beta \in (0, 1]$ is a parameter used to scale probabilities; and, $c, c' > 0$ are generic constants whose value may vary even within different lines of the same derivation. Let $\mathbf{e}_1, \ldots, \mathbf{e}_m$ be the standard basis vectors in $\mathbb{R}^m$. Let $\mathrm{A} \in \mathbb{R}^{m \times d}$ denote an arbitrary matrix which represents m points in $\mathbb{R}^d$. In general, we might represent a matrix such as A (roman, uppercase) by a set of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^d$ (bold, lowercase), so that $\mathrm{A}^{\mathrm{T}} = [\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_m]$; similarly, for a vector $\mathbf{y}$, $\mathbf{y}^{\mathrm{T}} = [y_1, \ldots, y_m]$. Note that $\mathbf{a}_t$ is the t^{th} row of A, which we may also refer to by $\mathrm{A}_{(t)}$; similarly, we may refer to the t^{th} column as $\mathrm{A}^{(t)}$. Let $\mathrm{rank}(\mathrm{A}) \leq \min\{m, d\}$ be the rank of A; typically $m \gg d$ and for concreteness, we will assume that $\mathrm{rank}(\mathrm{A}) = d$ (all the results easily generalize to $\mathrm{rank}(\mathrm{A}) < d$). For matrices, we will use the spectral norm, $\|\cdot\|$; on occasion, we will use the Frobenius norm, $\|\cdot\|_F$. For vectors, $\|\cdot\|_F = \|\cdot\|$ (the standard Euclidean norm). The stable, or "soft" rank, $\rho(\mathrm{A}) = \|\mathrm{A}\|_F^2/\|\mathrm{A}\|^2 \leq \mathrm{rank}(\mathrm{A})$.

The singular value decomposition (SVD) of A is

$$\mathrm{A} = \mathrm{U}_{\mathrm{A}}\mathrm{S}_{\mathrm{A}}\mathrm{V}_{\mathrm{A}}^{\mathrm{T}}.$$

where U_A is an $m \times d$ set of columns which are an orthonotmal basis for the column space in A; S_{A} is a $d \times d$ positive diagonal matrix of singular values, and V is a $d \times d$ orthogonal matrix. We refer to the singular values of A (the diagonal entries in S_{A}) by $\sigma_i(\mathrm{A})$. We will call a matrix with orthonormal columns an orthonormal matrix; an orthogonal matrix is a square orthonormal

matrix. In particular, $\mathrm{U_A^T U_A} = \mathrm{V_A^T V_A} = \mathrm{V_A V_A^T} = \mathrm{I}_{d\times d}$. It is possible to extend $\mathrm{U_A}$ to a full orthonormal basis of $\mathbb{R}^m$, $[\mathrm{U_A}, \mathrm{U_A^\perp}]$.

The SVD is important for a number of reasons. The projection of the columns of A onto the k left singular vectors with top k singular values gives the best rank-k approximation to A in the spectral and Frobenius norms. The solution to the linear regression problem is also intimately related to the SVD. In particular, consider the following minimization problem which is minimized at $\mathbf{w}^*$:

$$Z^* = \min_{\mathbf{w}} \|\mathrm{A}\mathbf{w} - \mathbf{y}\|^2.$$

It is known (Golub and van Loan, 1996) that $Z^* = \left\|\mathrm{U}_A^\perp(\mathrm{U_A^\perp})^\mathrm{T}\mathbf{y}\right\|^2$, and $\mathbf{w}^* = \mathrm{V_A S_A^{-1} U_A^T}\mathbf{y}$.

Row-Sampling Matrices Our focus is algorithms based on row-sampling. A *row-sampling matrix* $\mathrm{Q} \in \mathbb{R}^{r\times m}$ samples r rows of A to form $\tilde{\mathrm{A}} = \mathrm{QA}$:

$$\mathrm{Q} = \begin{bmatrix} \mathbf{r}_1^\mathrm{T} \\ \vdots \\ \mathbf{r}_r^\mathrm{T} \end{bmatrix}, \qquad \tilde{\mathrm{A}} = \mathrm{QA} = \begin{bmatrix} \mathbf{r}_1^\mathrm{T}\mathrm{A} \\ \vdots \\ \mathbf{r}_r^\mathrm{T}\mathrm{A} \end{bmatrix} = \begin{bmatrix} \lambda_{t_1}\mathbf{a}_{t_1}^\mathrm{T} \\ \vdots \\ \lambda_{t_r}\mathbf{a}_{t_r}^\mathrm{T} \end{bmatrix},$$

where $\mathbf{r}_j = \lambda_{t_j}\mathbf{e}_{t_j}$; it is easy to verify that the row $\mathbf{r}_j^\mathrm{T}\mathrm{A}$ samples the t_j^{th} row of A and rescales it. We are interested in random sampling matrices where each $\mathbf{r}_j$ is i.i.d. according to some distribution. Define a set of sampling probabilities $p_1, \ldots, p_m$, with $p_i \geq 0$ and $\sum_{i=1}^m p_i = 1$; then $\mathbf{r}_j = \mathbf{e}_t/\sqrt{rp_t}$ with probability p_t. Note that the scaling is also related to the sampling probabilities in all the algorithms we consider. We can write $\mathrm{Q^TQ}$ as the sum of r independently sampled matrices,

$$\mathrm{Q^TQ} = \frac{1}{r}\sum_{j=1}^{r} \mathbf{r}_j\mathbf{r}_j^\mathrm{T}$$

where $\mathbf{r}_j\mathbf{r}_j^\mathrm{T}$ is a diagonal matrix with only one non-zero diagonal entry; the t^{th} diagonal entry is equal to $1/p_t$ with probability p_t. Thus, by construction, for any set of non-zero sampling probabilities, $\mathbb{E}[\mathbf{r}_j\mathbf{r}_j^\mathrm{T}] = \mathrm{I}_{m\times m}$. Since we are averaging r independent copies, it is reasonable to expect a concentration around the mean, with respect to r, and so in some sense, $\mathrm{Q^TQ}$ essentially behaves like the identity.

1.2 Statement of Results

The two main results relate to how orthonormal subspaces behave with respect to the row-sampling. These are discussed more thoroughly in Section 3, but we state them here summarily.

Theorem 1 (Symmetric Orthonormal Subspace Sampling)**.** *Let* $\mathrm{U} \in \mathbb{R}^{m\times d}$ *be orthonormal, and* $\mathrm{S} \in \mathbb{R}^{d\times d}$ *be positive diagonal. Assume the row-sampling probabilities* p_t *satisfy*

$$p_t \geq \beta\frac{\mathbf{u}_t^\mathrm{T}\mathrm{S}^2\mathbf{u}_t}{\mathrm{trace}(\mathrm{S}^2)}.$$

Then, if $r \geq (4\rho(\mathrm{S})/\beta\epsilon^2)\ln\frac{2d}{\delta}$*, with probability at least* $1-\delta$*,*

$$\|\mathrm{S}^2 - \mathrm{SU^TQ^TQUS}\| \leq \epsilon\|\mathrm{S}\|^2$$

We also have an asymmetric version of Theorem 1, which is actually obtained through an application of Theorem 1 to a composite matrix.

Theorem 2 (Asymmetric Orthonormal Subspace Sampling). *Let* $\mathrm{W} \in \mathbb{R}^{m \times d_1}$, $\mathrm{V} \in \mathbb{R}^{m \times d_2}$ *be orthonormal, and let* $\mathrm{S}_1 \in \mathbb{R}^{d_1 \times d_1}$ *and* $\mathrm{S}_2 \in \mathbb{R}^{d_2 \times d_2}$ *be two positive diagonal matrices; let* $\rho_i = \rho(\mathrm{S}_i)$. *Consider row sampling probabilities*

$$p_t \geq \beta \frac{\frac{1}{\|\mathrm{S}_1\|^2}\mathbf{w}_t^{\mathrm{T}}\mathrm{S}_1^2\mathbf{w}_t + \frac{1}{\|\mathrm{S}_2\|^2}\mathbf{v}_t^{\mathrm{T}}\mathrm{S}_2^2\mathbf{v}_t}{\rho_1 + \rho_2}.$$

If $r \geq (8(\rho_1+\rho_2)/\beta\epsilon^2)\ln\frac{2(d_1+d_2)}{\delta}$, *then with probability at least* $1-\delta$,

$$\|\mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{V}\mathrm{S}_2 - \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2\| \leq \epsilon\|\mathrm{S}_1\|\|\mathrm{S}_2\|$$

We note that these row sampling probabilities are not the usual product row sampling probabilities one uses for matrix multiplication as in Drineas *et al.* (2006a). Computing the probabilities requires knowledge of the spectral norms of S_i. Here, S_i are given diagonal matrices, so it is easy to compute $\|\mathrm{S}_i\|$. In the application of these results to matrix multiplication, the spectral norm of the input matrices will appear. We will show how to handle this issue later. As a byproduct, we will give an efficient algorithm to obtain a relative error approximation to $\|\mathrm{A}\|$ based on row sampling and the power-iteration, which improves upon Woolfe *et al.* (2008); Kuczyński and Woźniakowski (1989).

We now give some applications of these orthonormal subspace sampling results.

Theorem 3 (Matrix Multiplication in Spectral Norm). *Let* $\mathrm{A} \in \mathbb{R}^{m \times d_1}$ *and* $\mathrm{B} \in \mathbb{R}^{m \times d_2}$ *have rescaled rows* $\hat{\mathbf{a}}_t = \mathbf{a}_t/\|\mathrm{A}\|$ *and* $\hat{\mathbf{b}}_t = \mathbf{b}_t/\|\mathrm{B}\|$ *respectively. Let* ρ_A *(resp.* ρ_B*) be the stable rank of* A *(resp.* B*). Obtain a sampling matrix* $\mathrm{Q} \in \mathbb{R}^{r \times m}$ *using row-sampling probabilities* p_t *satisfying*

$$p_t \geq \beta \frac{\hat{\mathbf{a}}_t^{\mathrm{T}}\hat{\mathbf{a}}_t + \hat{\mathbf{b}}_t^{\mathrm{T}}\hat{\mathbf{b}}_t}{\sum_{t=1}^m \hat{\mathbf{a}}_t^{\mathrm{T}}\hat{\mathbf{a}}_t + \hat{\mathbf{b}}_t^{\mathrm{T}}\hat{\mathbf{b}}_t} = \beta \frac{\hat{\mathbf{a}}_t^{\mathrm{T}}\hat{\mathbf{a}}_t + \hat{\mathbf{b}}_t^{\mathrm{T}}\hat{\mathbf{b}}_t}{\rho_{\mathrm{A}} + \rho_{\mathrm{B}}}.$$

Then, if $r \geq \frac{8(\rho_{\mathrm{A}}+\rho_{\mathrm{B}})}{\beta\epsilon^2}\ln\frac{2(d_1+d_2)}{\delta}$, *with probability at least* $1-\delta$,

$$\|\mathrm{A}^{\mathrm{T}}\mathrm{B} - \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{B}}\| \leq \epsilon\|\mathrm{A}\|\|\mathrm{B}\|.$$

The sampling probabilities depend on $\|\mathrm{A}\|^2$ and $\|\mathrm{B}\|^2$. It is possible to get a constant factor approximation to $\|\mathrm{A}\|^2$ (and similarly $\|\mathrm{B}\|^2$) with high probability. We summarize the idea here, the details are given in Section 7, Theorem 25. First sample $\tilde{\mathrm{A}} = \mathrm{Q}\mathrm{A}$ according to probabilities $p_t = \mathbf{a}_t^2/\|\mathrm{A}\|_F^2$. These probabilities are easy to compute in $O(md_1)$. By an application of the symmetric subspace sampling theorem (see Theorem 17), if $r \geq (4\rho_A/\epsilon^2)\ln\frac{2d_1}{\delta}$, then with probability at least $1-\delta$,

$$(1-\epsilon)\|\mathrm{A}\|^2 \leq \|\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\| \leq (1+\epsilon)\|\mathrm{A}\|^2.$$

We now run $\Omega(\ln\frac{d_1}{\delta})$ power iterations starting from a random isotropic vector to estimate the spectral norm of $\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}$. The efficiency is $O(md_1 + \rho_A d_1/\epsilon^2 \ln^2(\frac{d_1}{\delta}))$.

Theorem 4 (Sparse Row-Based Matrix Reconstruction). *Let* A *have the SVD representation* $\mathrm{A} = \mathrm{U}\mathrm{S}\mathrm{V}^{\mathrm{T}}$, *and consider row-sampling probabilities* p_t *satisfying* $p_t \geq \frac{\beta}{d}\mathbf{u}_t^{\mathrm{T}}\mathbf{u}_t$. *Then, if* $r \geq (4(d-\beta)/\beta\epsilon^2)\ln\frac{2d}{\delta}$, *with probability at least* $1-\delta$,

$$\|\mathrm{A} - \mathrm{A}\tilde{\Pi}_k\| \leq \left(\frac{1+\epsilon}{1-\epsilon}\right)^{1/2}\|\mathrm{A} - \mathrm{A}_k\|,$$

for $k = 1, \ldots, d$, *where* $\tilde{\Pi}_k$ *projects onto the top* k *right singular vectors of* $\tilde{\mathrm{A}}$.

It is possible to obtain relative approximations to the sampling probabilities according to the rows of the left singluar matrix (the leverage scores), but that goes beyond the scope of this work Magdon-Ismail (2010); Drineas *et al.* (2010)

Theorem 5 (Relative Error ℓ_2 Regression)**.** *Let* $\mathrm{A} \in \mathbb{R}^{m\times d}$ *have the SVD representation* $\mathrm{A} = \mathrm{U}\mathrm{S}\mathrm{V}^{\mathrm{T}}$*, and let* $\mathbf{y} \in \mathbb{R}^m$*. Let* $\mathbf{x}^* = \mathrm{A}^+\mathbf{y}$ *be the optimal regression with residual* $\boldsymbol{\epsilon} = \mathbf{y} - \mathrm{A}\mathbf{x}^* = \mathbf{y} - \mathrm{A}\mathrm{A}^+\mathbf{y}$*. Assume the sampling probabilities* p_t *satisfy*

$$p_t \geq \beta\left(\frac{\mathbf{u}_t^2}{d} + \frac{(\mathbf{u}_t^2 + \frac{\epsilon_t^2}{\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}})}{d+1} + \frac{\epsilon_t^2}{\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}}\right)$$

For $r \geq (8(d+1)/\beta\epsilon^2)\ln\frac{2(d+1)}{\delta}$*, let* $\hat{\mathbf{x}} = (\mathrm{Q}\mathrm{A})^+\mathrm{Q}\mathbf{y}$ *be the approximate regression. Then, with probability at least* $1-3\delta$*,*

$$\|\mathrm{A}\hat{\mathbf{x}} - \mathbf{y}\| \leq \left(1 + \epsilon + \epsilon\sqrt{\frac{1+\epsilon}{1-\epsilon}}\right)\|\mathrm{A}\mathbf{x}^* - \mathbf{y}\|.$$

In addition to sampling according to $\mathbf{u}_t^2$ we also need the residual vector $\boldsymbol{\epsilon} = \mathbf{y} - \mathrm{A}\mathrm{A}^+\mathbf{y}$. Unfortunately, we have not yet found an efficient way to get a good approximation (in some form of relative error) to this residual vector.

1.3 Paper Outline

Next we describe some probabistic tail inequalities which will be useful. We continue with the sampling lemmas for orthonormal matrices, followed by the applications to matrix multiplication, matrix reconstruction and ℓ_2-regression. Finally, we discuss the algorithm for approximating the spectral norm based on sampling and the power iteration.

2 Probabilistic Tail Inequalities

Since all our arguments involve high probability results, our main bounding tools will be probability tail inequalities. First, let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$ and $|\mathbf{X}_i| \leq \gamma$; let $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$. Chernoff, and later Hoeffding gave the bound

Theorem 6 (Chernoff (1952); Hoeffding (1963))**.** $\mathbb{P}[|Z_n| > \epsilon] \leq 2e^{-n\epsilon^2/2\gamma^2}$.

If in addition one can bound the variance, $\mathbb{E}[X_i^2] \leq s^2$, then we have Bernstein's bound:

Theorem 7 (Bernstein (1924))**.** $\mathbb{P}[|Z_n| \geq \epsilon] \leq 2e^{-n\epsilon^2/(2s^2+2\gamma\epsilon/3)}$.

Note that when $\epsilon \leq 3s^2/\gamma$, we can simplify the Bernstein bound to $\mathbb{P}[|Z_n| \geq \epsilon] \leq 2e^{-n\epsilon^2/4s^2}$, which is considerably simpler and only involves the variance. The non-commutative versions of these bounds, which extend these inequalities to matrix valued random variables can also be deduced. Let $\mathrm{X}_1, \ldots, \mathrm{X}_n$ be independent copies of a symmetric random matrix X, with $\mathbb{E}[\mathrm{X}] = \mathbf{0}$, and suppose that $\|\mathrm{X}\|_2 \leq \gamma$; let $\mathrm{Z}_n = \frac{1}{n}\sum_{i=1}^n \mathrm{X}_i$. Ahlswede and Winter (2002) gave the fundamental extension of the exponentiation trick for computing Chernoff bounds of scalar random variables to matrix valued random variables (for a simplified proof, see Wigderson and Xiao (2008)):

$$\mathbb{P}\left[\|\mathrm{Z}_n\|_2 > \epsilon\right] \leq \inf_t 2d e^{-n\epsilon t/\gamma}\|\,\mathbb{E}\left[e^{t\mathrm{X}/\gamma}\right]\|_2^n. \tag{1}$$

By standard optimization of this bound, one readily obtains the non-commutative tail inequality

Theorem 8 (Ahlswede and Winter (2002))**.** $\mathbb{P}[\|\mathrm{Z}_n\|_2 > \epsilon] \leq 2de^{-n\epsilon^2/4\gamma^2}$.

Proof. The statement is trivial if $\epsilon \geq \gamma$, so assume $\epsilon < \gamma$. The lemma follows from (1) and the following sequence after setting $t = \epsilon/2\gamma \leq \frac{1}{2}$:

$$\|\,\mathbb{E}\,[e^{t\mathrm{X}/\gamma}]\|_2 \overset{(a)}{\leq} 1 + \sum_{\ell=2}^{\infty} \frac{t^\ell}{\ell!}\,\mathbb{E}\,[\|(\mathrm{X}/\gamma)^\ell\|_2] \overset{(b)}{\leq} 1 + t^2 \leq e^{t^2}, \tag{2}$$

where (a) follows from $\mathbb{E}[\mathrm{X}] = 0$, the triangle inequality and $\|\,\mathbb{E}\,[\cdot]\|_2 \leq \mathbb{E}[\|\cdot\|_2]$; (b) follows because $\|\mathrm{X}/\gamma\|_2 \leq 1$ and $t \leq \frac{1}{2}$. ∎

(We have stated a simplified version of the bound, without taking care to optimize the constants.) As mentioned in Gross *et al.* (2009), one can obtain a non-commuting version of Bernstein's inequality in a similar fashion using (1). Assume that $\|\,\mathbb{E}\,\mathrm{X}^{\mathrm{T}}\mathrm{X}\|_2 \leq s^2$. By adapting the standard Bernstein bounding argument to matrices, we have

Lemma 9. *For symmetric* X, $\|\,\mathbb{E}\,[e^{t\mathrm{X}/\gamma}]\|_2 \leq \exp\left(\frac{s^2}{\gamma^2}(e^t - 1 - t)\right)$.

Proof. As in (2), but using submultiplicativity, we first bound $\|\,\mathbb{E}\,[\mathrm{X}^\ell]\|_2 \leq s^2\gamma^{\ell-2}$:

$$\begin{aligned}
\|\,\mathbb{E}\,[\mathrm{X}^\ell]\|_2 &= \max_{\|\mathbf{u}\|=1} \left\| \int d\mathrm{X}\; p(\mathrm{X})\mathrm{X}^\ell \mathbf{u} \right\| \\
&= \max_{\|\mathbf{u}\|=1} \left\| \int d\mathrm{X}\; p(\mathrm{X}) \frac{\|\mathrm{X}^{\ell-2}\mathbf{u}\|\mathrm{X}^2\mathrm{X}^{\ell-2}\mathbf{u}}{\|\mathrm{X}^{\ell-2}\mathbf{u}\|} \right\| \\
&\leq \gamma^{\ell-2} \max_{\|\mathbf{w}\|=1} \left\| \int d\mathrm{X}\; p(\mathrm{X})\mathrm{X}^2\mathbf{w} \right\| \\
&= \gamma^{\ell-2}\|\,\mathbb{E}\,[\mathrm{X}^2]\|_2 \leq s^2\gamma^{\ell-2}.
\end{aligned}$$

To conclude, we use the triangle inequality to bound as follows:

$$\|\,\mathbb{E}\,[e^{t\mathrm{X}/\gamma}]\|_2 = \left\| \mathrm{I} + \sum_{\ell=2}^{\infty} \frac{t^\ell}{\gamma^\ell \ell!}\,\mathbb{E}\,[\mathrm{X}^\ell] \right\|_2 \leq 1 + \frac{s^2}{\gamma^2}\sum_{\ell=2}^{\infty}\frac{t^\ell}{\ell!} = 1 + \frac{s^2}{\gamma^2}(e^t - 1 - t) \leq \exp\left(\tfrac{s^2}{\gamma^2}(e^t - 1 - t)\right).$$

∎

Using Lemma 9 in (1) with $t = \ln(1 + \epsilon\gamma/s^2)$, and using $(1+x)\ln(1+\frac{1}{x}) - 1 \geq \frac{1}{2x+2/3}$, we obtain the following result.

Theorem 10 (Non-commutative Bernstein)**.** $\mathbb{P}[\|\mathrm{Z}_n\|_2 > \epsilon] \leq 2de^{-n\epsilon^2/(2s^2+2\gamma\epsilon/3)}$.

Gross *et al.* (2009) gives a simpler version of this non-commutative Bernstein inequality. If $\mathrm{X} \in \mathbb{R}^{d_1\times d_2}$ is not symmetric, then by considering

$$\begin{bmatrix} \mathbf{0}_{d_1\times d_1} & \mathrm{X} \\ \mathrm{X}^{\mathrm{T}} & \mathbf{0}_{d_2\times d_2} \end{bmatrix},$$

one can get a non-symmetric verision of the non-commutative Chernoff and Bernstein bounds,

Theorem 11 (Recht (2009))**.** $\mathbb{P}[\|\mathrm{Z}_n\|_2 > \epsilon] \leq (d_1 + d_2)e^{-n\epsilon^2/(2s^2+2\gamma\epsilon/3)}$.

For most of our purposes, we will only need the symmetric version; again, if $\epsilon \leq 3s^2/\gamma$, then we have the much simpler bound $\mathbb{P}[\|\mathrm{Z}_n\|_2 > \epsilon] \leq 2de^{-n\epsilon^2/4s^2}$.

3 Orthonormal Sampling Lemmas

Let $\mathrm{U} \in \mathbb{R}^{m\times d}$ be an orthonormal matrix, and let $\mathrm{S} \in \mathbb{R}^{d\times d}$ be a diagonal matrix. We are interested in the product $\mathrm{US} \in \mathbb{R}^{m\times d}$; US is the matrix with columns $\mathrm{U}^{(i)}S_{ii}$. Without loss of generality, we can assume that S is positive by flipping the signs of the appropriate columns of U. The row-representation of U is $\mathrm{U}^{\mathrm{T}} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$; we consider the row sampling probabilities

$$p_t \geq \beta \frac{\mathbf{u}_t^{\mathrm{T}} \mathrm{S}^2 \mathbf{u}_t}{\mathrm{trace}(\mathrm{S}^2)}. \tag{3}$$

Since $\mathrm{U}^{\mathrm{T}}\mathrm{U} = \mathrm{I}_{d\times d}$, one can verify that $\mathrm{trace}(\mathrm{S}^2) = \sum_t \mathbf{u}_t^{\mathrm{T}}\mathrm{S}^2\mathbf{u}_t$ is the correct normalization.

Lemma 12 (Symmetric Subspace Sampling Lemma)**.**

$$\begin{aligned}\mathbb{P}[\|\mathrm{S}^2 - \mathrm{SU}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{QUS}\| > \epsilon\|\mathrm{S}\|^2] &\leq 2d \cdot \exp\left(\frac{-r\epsilon^2}{2(\rho/\beta - \kappa^{-4} + \epsilon(\rho/\beta - \kappa^{-2})/3)}\right), \\ &\leq 2d \cdot \exp\left(\frac{-r\beta\epsilon^2}{4\rho}\right),\end{aligned}$$

where ρ is the numerical (stable) rank of S, $\rho(\mathrm{S}) = \|\mathrm{S}\|_F^2/\|\mathrm{S}\|^2$, *and* $\kappa(\mathrm{S}) = \sigma_{\max}(\mathrm{S})/\sigma_{\min}(\mathrm{S})$ *is the condition number.*

Remarks. The stable rank $\rho \leq d$ measures the effective dimension of the matrix. The condition number $\kappa \geq 1$, hence the simpler version of the bound, which is valid for $\epsilon \leq 3$. It immediately follows that if $r \geq (4\rho/\beta\epsilon^2)\ln\frac{2d}{\delta}$, then with probability at least $1-\delta$,

$$\|\mathrm{S}^2 - \mathrm{SU}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{QUS}\| \leq \epsilon\|\mathrm{S}\|^2$$

An important special case is when $\mathrm{S} = \mathrm{I}_{d\times d}$, in which case $\rho = d$, $\kappa = 1$ and $\|\mathrm{S}\| = 1$.

Corollary 13. *For sampling probabilities $p_t \geq \frac{\beta}{d}\mathbf{u}_t^{\mathrm{T}}\mathbf{u}_t$,*

$$\mathbb{P}[\|\mathrm{I} - \mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{QU}\| > \epsilon] \leq 2d \cdot \exp\left(\frac{-\beta r\epsilon^2}{4(d-\beta)}\right),$$

Proof. (of Lemma 12) Note that $\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{QU} = \frac{1}{r}\sum_{i=1}^r \mathbf{u}_{t_i}\mathbf{u}_{t_i}^{\mathrm{T}}/p_{t_i}$, where $t_i \in [1,m]$ is chosen according to the probability p_{t_i}. It follows that

$$\mathrm{S}^2 - \mathrm{SU}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{QUS} = \frac{1}{r}\sum_{i=1}^r \mathrm{S}^2 - \frac{1}{p_{t_i}}\mathrm{S}\mathbf{u}_{t_i}\mathbf{u}_{t_i}^{\mathrm{T}}\mathrm{S} = \frac{1}{r}\sum_{i=1} \mathrm{X}_i,$$

where X_i are independent copies of a matrix-random variable $\mathrm{X} \sim \mathrm{S}^2 - \mathrm{S}\mathbf{u}\mathbf{u}^{\mathrm{T}}\mathrm{S}/p$. We prove the following three claims:

$$\begin{aligned}&(i)\ \ \mathbb{E}[\mathrm{X}] = \mathbf{0};\\ &(ii)\ \|\mathrm{X}\| \leq \|\mathrm{S}\|^2(\rho/\beta - \kappa^{-2});\\ &(ii)\ \|\mathbb{E}\,\mathrm{X}^{\mathrm{T}}\mathrm{X}\| \leq \|\mathrm{S}\|^4(\rho/\beta - \kappa^{-4}).\end{aligned}$$

The Lemma follows from the non-commutative Bernstein bound with ϵ replaced by $\epsilon\|\mathrm{S}\|^2$. To prove (i), note that $\mathbb{E}[\mathrm{X}] = \mathrm{S}^2 - \mathrm{S}\,\mathbb{E}\,[\mathbf{u}\mathbf{u}^{\mathrm{T}}/p]\mathrm{S} = \mathrm{S}^2 - \mathrm{S}\left(\sum_{t=1}^m \mathbf{u}_t\mathbf{u}_t^{\mathrm{T}}\right)\mathrm{S} = \mathbf{0}$, because $\sum_{t=1}^m \mathbf{u}_t\mathbf{u}_t^{\mathrm{T}} = \mathrm{U}^{\mathrm{T}}\mathrm{U} = \mathrm{I}_{d\times d}$.

To prove (ii), let $\mathbf{z}$ be an arbitrary unit vector and consider

$$\mathbf{z}^{\mathrm{T}}\mathrm{X}\mathbf{z} = \mathbf{z}^{\mathrm{T}}\mathrm{S}^2\mathbf{z} - \frac{1}{p}(\mathbf{z}^{\mathrm{T}}\mathrm{S}\mathbf{u})^2.$$

It follows that $\mathbf{z}^{\mathrm{T}}\mathrm{X}\mathbf{z} \leq \|\mathrm{S}\|^2$. To get a lower bound, we use $p \geq \beta\mathbf{u}^{\mathrm{T}}\mathrm{S}^2\mathbf{u}/\mathrm{trace}(\mathrm{S}^2)$:

$$\begin{aligned}
\mathbf{z}^{\mathrm{T}}\mathrm{X}\mathbf{z} &\geq \mathbf{z}^{\mathrm{T}}\mathrm{S}^2\mathbf{z} - \frac{\mathrm{trace}(\mathrm{S}^2)}{\beta}\frac{(\mathbf{z}^{\mathrm{T}}\mathrm{S}\mathbf{u})^2}{\mathbf{u}^{\mathrm{T}}\mathrm{S}^2\mathbf{u}}, \\
&\overset{(a)}{\geq} \|\mathrm{S}\|^2\left(\frac{\sigma^2_{\min}(\mathrm{S})}{\|\mathrm{S}\|^2} - \frac{\mathrm{trace}(\mathrm{S}^2)}{\beta\|S\|^2}\right), \\
&= \|\mathrm{S}\|^2\left(\frac{1}{\kappa^2} - \frac{\rho}{\beta}\right).
\end{aligned}$$

(a) follows because: by definition of $\sigma_{\min}$, the minimum of the first term is $\sigma^2_{\min}$; and, by Cauchy-Schwarz, $(\mathbf{z}^{\mathrm{T}}\mathrm{S}\mathbf{u})^2 \leq (\mathbf{z}^{\mathrm{T}}\mathbf{z})(\mathbf{u}^{\mathrm{T}}\mathrm{S}^2\mathbf{u})$. Since $\beta \leq 1$, $\rho/\beta - \kappa^{-2} \geq 1$ (for $d > 1$), and so $|\mathbf{z}^{\mathrm{T}}\mathrm{X}\mathbf{z}| \leq \|\mathrm{S}\|^2\left(\rho/\beta - \kappa^{-2}\right)$, from which (ii) follows.

To prove (iii), first note that

$$\begin{aligned}
\mathbb{E}[\mathrm{X}^{\mathrm{T}}\mathrm{X}] &= \mathrm{S}^4 - \mathrm{S}^3\,\mathbb{E}\left[\mathbf{u}\mathbf{u}^{\mathrm{T}}/p\right]\mathrm{S} - \mathrm{S}\,\mathbb{E}\left[\mathbf{u}\mathbf{u}^{\mathrm{T}}/p\right]\mathrm{S}^3 + \mathrm{S}\,\mathbb{E}\left[\mathbf{u}\mathbf{u}^{\mathrm{T}}\mathrm{S}^2\mathbf{u}\mathbf{u}^{\mathrm{T}}/p^2\right]\mathrm{S}, \\
&\overset{(a)}{=} \mathrm{S}\left(\sum_{t=1}^{m}\frac{1}{p_t}\mathbf{u}_t\mathbf{u}_t^{\mathrm{T}}\mathrm{S}^2\mathbf{u}_t\mathbf{u}_t^{\mathrm{T}}\right)\mathrm{S} - \mathrm{S}^4.
\end{aligned}$$

(a) follows because $\mathbb{E}[\mathbf{u}\mathbf{u}^{\mathrm{T}}/p] = \mathrm{I}$. Thus, for an arbitrary unit $\mathbf{z}$, we have

$$\begin{aligned}
\mathbf{z}^{\mathrm{T}}\,\mathbb{E}\left[\mathrm{X}^{\mathrm{T}}\mathrm{X}\right]\mathbf{z} &= \sum_{t=1}^{m}\frac{1}{p_t}(\mathbf{z}^{\mathrm{T}}\mathrm{S}\mathbf{u}_t\mathbf{u}_t^{\mathrm{T}}\mathrm{S}\mathbf{z})\mathbf{u}_t^{\mathrm{T}}\mathrm{S}^2\mathbf{u}_t - \mathbf{z}^{\mathrm{T}}\mathrm{S}^4\mathbf{z}^{\mathrm{T}}, \\
&\overset{(a)}{\leq} \frac{\mathrm{trace}(\mathrm{S}^2)}{\beta}\mathbf{z}^{\mathrm{T}}\mathrm{S}\left(\sum_{t=1}^{m}\mathbf{u}_t\mathbf{u}_t^{\mathrm{T}}\right)\mathrm{S}\mathbf{z} - \mathbf{z}^{\mathrm{T}}\mathrm{S}^4\mathbf{z}^{\mathrm{T}}, \\
&\overset{(b)}{=} \|\mathrm{S}\|^4\left(\frac{\mathrm{trace}(\mathrm{S}^2)}{\beta\|\mathrm{S}\|^2}\frac{\mathbf{z}^{\mathrm{T}}\mathrm{S}^2\mathbf{z}^{\mathrm{T}}}{\|\mathrm{S}\|^2} - \frac{\mathbf{z}^{\mathrm{T}}\mathrm{S}^4\mathbf{z}^{\mathrm{T}}}{\|\mathrm{S}\|^4}\right), \\
&\leq \|\mathrm{S}\|^4\left(\frac{\mathrm{trace}(\mathrm{S}^2)}{\beta\|\mathrm{S}\|^2} - \frac{\sigma^4_{\min}}{\|\mathrm{S}\|^4}\right).
\end{aligned}$$

(a) follows from $p_t \geq \beta\mathbf{u}_t^{\mathrm{T}}\mathrm{S}^2\mathbf{u}_t/\mathrm{trace}(\mathrm{S}^2)$; (b) follows from $\mathrm{U}^{\mathrm{T}}\mathrm{U} = \sum_{t=1}^{m}\mathbf{u}_t\mathbf{u}_t^{\mathrm{T}} = \mathrm{I}_{d\times d}$. Thus, $|\mathbf{z}^{\mathrm{T}}\,\mathbb{E}\left[\mathrm{X}^{\mathrm{T}}\mathrm{X}\right]\mathbf{z}| \leq \|\mathrm{S}\|^4(\rho/\beta - \kappa^{-4})$, from which (iii) follows. ■

For the general case, consider two orthonormal matrices $\mathrm{W} \in \mathbb{R}^{m\times d_1}$, $\mathrm{V} \in \mathbb{R}^{m\times d_2}$, and two positive diagonal matrices $\mathrm{S}_1 \in \mathbb{R}^{d_1\times d_1}$ and $\mathrm{S}_2 \in \mathbb{R}^{d_2\times d_2}$. We consider the product $\mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{V}\mathrm{S}_2$, which is approximated by the sampled product $\mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2$. Consider the sampling probabilities

$$p_t \geq \beta\frac{(\mathbf{u}_t^{\mathrm{T}}\mathrm{S}_1^2\mathbf{u}_t)^{1/2}(\mathbf{v}_t^{\mathrm{T}}\mathrm{S}_2^2\mathbf{v}_t)^{1/2}}{\sum_{t=1}^{m}(\mathbf{u}_t^{\mathrm{T}}\mathrm{S}_1^2\mathbf{u}_t)^{1/2}(\mathbf{v}_t^{\mathrm{T}}\mathrm{S}_2^2\mathbf{v}_t)^{1/2}} \geq \beta\frac{(\mathbf{u}_t^{\mathrm{T}}\mathrm{S}_1^2\mathbf{u}_t)^{1/2}(\mathbf{v}_t^{\mathrm{T}}\mathrm{S}_2^2\mathbf{v}_t)^{1/2}}{\sqrt{\mathrm{trace}(\mathrm{S}_1^2)\mathrm{trace}(\mathrm{S}_2^2)}},$$

where the last inequality follows from Cauchy-Schwarz. Since $\|\mathrm{A}\|_F = \sqrt{\rho(\mathrm{A})}\|\mathrm{A}\| \geq \|\mathrm{A}\|$, any bound for the Frobenius norm can be converted into a bound for the spectral norm. Using the

Frobenius norm bounds in Drineas *et al.* (2006a) (using a simplified form for the bound), one immediately has:

$$\mathbb{P}\left[\|\mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{V}\mathrm{S}_2 - \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2\| > \epsilon\|\mathrm{S}_1\|\|\mathrm{S}_2\|\right] \leq \exp\left(\frac{-r\beta^2\epsilon^2}{16\rho_1\rho_2}\right), \tag{4}$$

where $\rho_1 = \rho(\mathrm{S}_1)$ and $\rho_2 = \rho(\mathrm{S}_2)$. Alternatively, if $r \geq (16\rho_1\rho_2/\beta^2\epsilon^2)\ln\frac{1}{\delta}$, then

$$\|\mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{V}\mathrm{S}_2 - \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2\| \leq \epsilon\|\mathrm{S}_1\|\|\mathrm{S}_2\|.$$

The dependence on the stable ranks and β is quadratic. Applying this bound to the situation in Lemma 12 would give an inferior bound. The intuition behind the improvement is that the sampling is isotropic, and so will not favor any particular direction. One can therefore guess that all the singular values are approximately equal and so the Frobenius norm bound on the spectral norm will be loose by a factor of $\sqrt{\rho}$; and, indeed this is what comes out in the closer analysis. As a application of Lemma 12, we can get a better result for the asymmetric case.

Lemma 14. *Let* $\mathrm{W} \in \mathbb{R}^{m\times d_1}$, $\mathrm{V} \in \mathbb{R}^{m\times d_2}$ *be orthonormal, and let* $\mathrm{S}_1 \in \mathbb{R}^{d_1\times d_1}$ *and* $\mathrm{S}_2 \in \mathbb{R}^{d_2\times d_2}$ *be two positive diagonal matrices. Consider row sampling probabilities*

$$p_t \geq \beta\frac{\frac{1}{\|\mathrm{S}_1\|^2}\mathbf{w}_t^{\mathrm{T}}\mathrm{S}_1^2\mathbf{w}_t + \frac{1}{\|\mathrm{S}_2\|^2}\mathbf{v}_t^{\mathrm{T}}\mathrm{S}_2^2\mathbf{v}_t}{\rho_1+\rho_2}.$$

If $r \geq (8(\rho_1+\rho_2)/\beta\epsilon^2)\ln\frac{2(d_1+d_2)}{\delta}$, *then with probability at least* $1-\delta$,

$$\|\mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{V}\mathrm{S}_2 - \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2\| \leq \epsilon\|\mathrm{S}_1\|\|\mathrm{S}_2\|$$

For the special case that $\mathrm{S}_1 = \mathrm{I}_{d_1\times d_1}$ and $\mathrm{S}_2 = \mathrm{I}_{d_2\times d_2}$, the sampling probabilities simplify to

$$p_t \geq \beta\frac{\mathbf{w}_t^{\mathrm{T}}\mathbf{w}_t + \mathbf{v}_t^{\mathrm{T}}\mathbf{v}_t}{d_1+d_2},$$

Corollary 15. *If* $r \geq (8(d_1+d_2)/\beta\epsilon^2)\ln\frac{2(d_1+d_2)}{\delta}$, *then with probability at least* $1-\delta$,

$$\|\mathrm{W}^{\mathrm{T}}\mathrm{V} - \mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\| \leq \epsilon.$$

Proof. (of Lemma 14) By homogeneity, we can without loss of generality assume that $\|\mathrm{S}_1\| = \|\mathrm{S}_2\| = 1$, and let[1] $\mathrm{Z} = [\mathrm{W}\mathrm{S}_1\ \mathrm{V}\mathrm{S}_2]$. An elementary lemma which we will find useful is

Lemma 16. *For any matrix* $\mathrm{A} = [\mathrm{A}_1\ \mathrm{A}_2]$,

$$\max\{\|\mathrm{A}_1\|, \|\mathrm{A}_2\|\} \leq \|\mathrm{A}\| \leq \sqrt{\|\mathrm{A}_1\|^2 + \|\mathrm{A}_2\|^2}.$$

The left inequality is saturated when A_1 and A_2 are orthogonal ($\mathrm{A}_1^{\mathrm{T}}\mathrm{A}_2 = \mathbf{0}$), and the right inequality is saturated when $\mathrm{A}_1 = \mathrm{A}_2$. By repeatedly applying Lemma 16 one can see that $\|\mathrm{A}\|$ is at least the spectral norm of any submatrix. Introduce the SVD of Z,

$$\mathrm{Z} = [\mathrm{W}\mathrm{S}_1\ \mathrm{V}\mathrm{S}_2] = \mathrm{U}\mathrm{S}\mathrm{V}_{\mathrm{Z}}^{\mathrm{T}}.$$

[1]The general case would have been $\mathrm{Z} = \left[\frac{1}{\|\mathrm{S}_1\|}\mathrm{W}\mathrm{S}_1\ \frac{1}{\|\mathrm{S}_2\|}\mathrm{V}\mathrm{S}_2\right]$.

We now use the row sampling probabilities according to US from (3),

$$p_t \geq \beta \frac{\mathbf{u}_t^{\mathrm{T}} \mathrm{S}^2 \mathbf{u}_t}{\mathrm{trace}(\mathrm{S}^2)}.$$

We may interpret the sampling probabilities as follows. Let $\mathbf{z}_t$ be a row of Z, the concatenation of two rows in WS_1 and VS_2: $\mathbf{z}_t^{\mathrm{T}} = [\mathbf{w}_t^{\mathrm{T}} \mathrm{S}_1 \ \mathbf{v}_t^{\mathrm{T}} \mathrm{S}_2]$. We also have that $\mathbf{z}_t^{\mathrm{T}} = \mathbf{u}_t^{\mathrm{T}} \mathrm{S} \mathrm{V}_{\mathrm{Z}}^{\mathrm{T}}$. Hence,

$$\mathbf{u}_t^{\mathrm{T}} \mathrm{S}^2 \mathbf{u}_t = \mathbf{u}_t^{\mathrm{T}} \mathrm{S} \mathrm{V}_{\mathrm{Z}}^{\mathrm{T}} \mathrm{V}_{\mathrm{Z}} \mathrm{S} \mathbf{u}_t = \mathbf{z}_t^{\mathrm{T}} \mathbf{z}_t = \mathbf{w}_t^{\mathrm{T}} \mathrm{S}_1^2 \mathbf{w}_t + \mathbf{v}_t^{\mathrm{T}} \mathrm{S}_2^2 \mathbf{v}_t.$$

These are exactly the probabilities as claimed in the statement of the lemma (modulo the rescaling).

Applying Lemma 12: if $r \geq (4\rho/\beta\epsilon^2) \ln \frac{2 \cdot \mathrm{rank(U)}}{\delta}$, then with probability at least $1 - \delta$,

$$\|\mathrm{S}^2 - \mathrm{S}\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\mathrm{S}\| \leq \epsilon \|\mathrm{S}\|^2 \leq \epsilon \sqrt{\|\mathrm{S}_1\|^2 + \|\mathrm{S}_2\|^2} = \epsilon\sqrt{2},$$

where the second inequality follows from Lemma 16. Since ZV = US,

$$\|\mathrm{Z}^{\mathrm{T}}\mathrm{Z} - \mathrm{Z}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{Z}\| = \|\mathrm{S}^2 - \mathrm{S}\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\mathrm{S}\|.$$

Further, by the construction of Z,

$$\mathrm{Z}^{\mathrm{T}}\mathrm{Z} - \mathrm{Z}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{Z} = \begin{bmatrix} \mathrm{S}_1^2 - \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{W}\mathrm{S}_1 & \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{V}\mathrm{S}_2 - \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2 \\ \mathrm{S}_2\mathrm{V}^{\mathrm{T}}\mathrm{W}\mathrm{S}_1 - \mathrm{S}_2\mathrm{V}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{W}\mathrm{S}_1 & \mathrm{S}_2^2 - \mathrm{S}_2\mathrm{V}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2 \end{bmatrix}.$$

By Lemma 16, $\|\mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{V}\mathrm{S}_2 - \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2\| \leq \|\mathrm{Z}^{\mathrm{T}}\mathrm{Z} - \mathrm{Z}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{Z}\|$, and so:

$$\|\mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{V}\mathrm{S}_2 - \mathrm{S}_1\mathrm{W}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{V}\mathrm{S}_2\| \leq \epsilon\sqrt{2}.$$

Observe that $\mathrm{trace}(\mathrm{S}^2) = \|\mathrm{Z}\|_F^2 = \mathrm{trace}(\mathrm{S}_1^2) + \mathrm{trace}(\mathrm{S}_2^2)$; further, since $\|\mathrm{S}\| \geq \max\{\|\mathrm{S}_1\|, \|\mathrm{S}_2\|\}$, we have that

$$\rho(\mathrm{S}) = \frac{\mathrm{trace}(\mathrm{S}^2)}{\|\mathrm{S}\|^2} = \frac{\mathrm{trace}(\mathrm{S}_1^2) + \mathrm{trace}(\mathrm{S}_2^2)}{\|\mathrm{S}\|^2} \leq \frac{\mathrm{trace}(\mathrm{S}_1^2)}{\|\mathrm{S}_1\|^2} + \frac{\mathrm{trace}(\mathrm{S}_2^2)}{\|\mathrm{S}_2\|^2} = \rho_1 + \rho_2.$$

Since $\mathrm{rank}(\mathrm{U}) \leq d_1 + d_2$, it suffices that $r \geq (4(\rho_1 + \rho_2)/\beta\epsilon^2) \ln \frac{2(d_1+d_2)}{\delta}$ to obtain error $\epsilon\sqrt{2}$; after rescaling $\epsilon' = \epsilon\sqrt{2}$, we have the result. ■

4 Sampling for Matrix Multiplication

We obtain results for matrix multiplication directly from Lemmas 12 and 14. First we consider the symmetric case, then the asymmetric case. Let $\mathrm{A} \in \mathbb{R}^{m \times d_1}$ and $\mathrm{B} \in \mathbb{R}^{m \times d_2}$. We are interested in conditions on the sampling matrix $\mathrm{Q} \in \mathbb{R}^{r \times m}$ such that $\mathrm{A}^{\mathrm{T}}\mathrm{A} \approx \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}$ and $\mathrm{A}^{\mathrm{T}}\mathrm{B} \approx \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{B}}$, where $\tilde{\mathrm{A}} = \mathrm{Q}\mathrm{A}$ and $\tilde{\mathrm{B}} = \mathrm{Q}\mathrm{B}$. Using the SVD of A,

$$\begin{aligned} \|\mathrm{A}^{\mathrm{T}}\mathrm{A} - \mathrm{A}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{A}\| &= \|\mathrm{V}_{\mathrm{A}}\mathrm{S}_{\mathrm{A}}\mathrm{U}_{\mathrm{A}}^{\mathrm{T}}\mathrm{U}_{\mathrm{A}}\mathrm{S}_{\mathrm{A}}\mathrm{V}_{\mathrm{A}}^{\mathrm{T}} - \mathrm{V}_{\mathrm{A}}\mathrm{S}_{\mathrm{A}}\mathrm{U}_{\mathrm{A}}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}_{\mathrm{A}}\mathrm{S}_{\mathrm{A}}\mathrm{V}_{\mathrm{A}}^{\mathrm{T}}\|, \\ &= \|\mathrm{S}_{\mathrm{A}}^2 - \mathrm{S}_{\mathrm{A}}\mathrm{U}_{\mathrm{A}}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}_{\mathrm{A}}\mathrm{S}_{\mathrm{A}}\|. \end{aligned}$$

We may now directly apply Lemma 12, with respect to the appropriate sampling probabilities. One can verify that the sampling probabilities in Lemma 12 are proportional to the squared norms of the rows of A.

Theorem 17. *Let* $\mathrm{A} \in \mathbb{R}^{m \times d_1}$ *have rows* $\mathbf{a}_t$ *Obtain a sampling matrix* $\mathrm{Q} \in \mathbb{R}^{r \times m}$ *using row-sampling probabilities*

$$p_t \geq \beta \frac{\mathbf{a}_t^{\mathrm{T}} \hat{\mathbf{a}}_t}{\|\mathrm{A}\|_F^2}.$$

Then, if $r \geq \frac{4\rho_{\mathrm{A}}}{\beta\epsilon^2} \ln \frac{2d_1}{\delta}$*, with probability at least* $1 - \delta$*,*

$$\|\mathrm{A}^{\mathrm{T}}\mathrm{A} - \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\| \leq \epsilon \|\mathrm{A}\|^2.$$

Similarly, using the SVDs of A and B,

$$\begin{aligned} \|\mathrm{A}^{\mathrm{T}}\mathrm{B} - \mathrm{A}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{B}\| &= \|\mathrm{V}_{\mathrm{A}}\mathrm{S}_{\mathrm{A}}\mathrm{U}_{\mathrm{A}}^{\mathrm{T}}\mathrm{U}_{\mathrm{B}}\mathrm{S}_{\mathrm{B}}\mathrm{V}_{\mathrm{B}}^{\mathrm{T}} - \mathrm{V}_{\mathrm{A}}\mathrm{S}_{\mathrm{A}}\mathrm{U}_{\mathrm{A}}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}_{\mathrm{B}}\mathrm{S}_{\mathrm{B}}\mathrm{V}_{\mathrm{B}}^{\mathrm{T}}\|, \\ &= \|\mathrm{S}_{\mathrm{A}}\mathrm{U}_{\mathrm{A}}^{\mathrm{T}}\mathrm{U}_{\mathrm{B}}\mathrm{S}_{\mathrm{B}} - \mathrm{S}_{\mathrm{A}}\mathrm{U}_{\mathrm{A}}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}_{\mathrm{B}}\mathrm{S}_{\mathrm{B}}\|. \end{aligned}$$

We may now directly apply Lemma 14, with respect to the appropriate sampling probabilities. One can verify that the sampling probabilities in Lemma 14 are proportional to the sum of the rescaled squared norms of the rows of A and B.

Theorem 18. *Let* $\mathrm{A} \in \mathbb{R}^{m \times d_1}$ *and* $\mathrm{B} \in \mathbb{R}^{m \times d_2}$*, have rescaled rows* $\hat{\mathbf{a}}_t = \mathbf{a}_t / \|\mathrm{A}\|$ *and* $\hat{\mathbf{b}}_t = \mathbf{b}_t / \|\mathrm{B}\|$ *respectively. Obtain a sampling matrix* $\mathrm{Q} \in \mathbb{R}^{r \times m}$ *using row-sampling probabilities*

$$p_t \geq \beta \frac{\hat{\mathbf{a}}_t^{\mathrm{T}}\hat{\mathbf{a}}_t + \hat{\mathbf{b}}_t^{\mathrm{T}}\hat{\mathbf{b}}_t}{\sum_{t=1}^{m} \hat{\mathbf{a}}_t^{\mathrm{T}}\hat{\mathbf{a}}_t + \hat{\mathbf{b}}_t^{\mathrm{T}}\hat{\mathbf{b}}_t} = \beta \frac{\hat{\mathbf{a}}_t^{\mathrm{T}}\hat{\mathbf{a}}_t + \hat{\mathbf{b}}_t^{\mathrm{T}}\hat{\mathbf{b}}_t}{\rho_{\mathrm{A}} + \rho_{\mathrm{B}}}.$$

Then, if $r \geq \frac{8(\rho_{\mathrm{A}} + \rho_{\mathrm{B}})}{\beta\epsilon^2} \ln \frac{2(d_1 + d_2)}{\delta}$*, with probability at least* $1 - \delta$*,*

$$\|\mathrm{A}^{\mathrm{T}}\mathrm{B} - \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{B}}\| \leq \epsilon \|\mathrm{A}\|\|\mathrm{B}\|.$$

5 Sparse Row Based Matrix Representation

Given a matrix $\mathrm{A} = \mathrm{U}\mathrm{S}\mathrm{V}^{\mathrm{T}} \in \mathbb{R}^{m \times d}$, the top k singular vectors, corresponding to the top k singular values give the best rank k reconstruction of A. Specifically, let $\mathrm{A}_k = \mathrm{U}_k\mathrm{S}_k\mathrm{V}_k^{\mathrm{T}}$, where $\mathrm{U}_k \in \mathbb{R}^{m \times k}$, $\mathrm{S}_k \in \mathbb{R}^{k \times k}$ and $\mathrm{V}_k \in \mathbb{R}^{d \times k}$; U_k and V_k correspond to the top-k left and right singular vectors. Then, $\|\mathrm{A} - \mathrm{A}_k\| \leq \|\mathrm{A} - \mathrm{X}\|$ where $\mathrm{X} \in \mathbb{R}^{m \times d}$ ranges over all rank-k matrices. As usual, let $\tilde{\mathrm{A}} = \mathrm{Q}\mathrm{A}$ be the sampled, rescaled rows of A, with $\tilde{\mathrm{A}} = \tilde{\mathrm{U}}\tilde{\mathrm{S}}\tilde{\mathrm{V}}^{\mathrm{T}}$, and consider the top-$k$ right singular vectors $\tilde{\mathrm{V}}_k$. Let $\tilde{\Pi}_k$ be the projection onto this top-k right singular space, and consider the rank k approximation to A obtained by projecting onto this space: $\tilde{\mathrm{A}}_k = \mathrm{A}\tilde{\Pi}_k$. The following lemma is useful for showing that $\tilde{\mathrm{A}}_k$ is almost (up to additive error) as good an approximation to A as one can get.

Lemma 19 (Drineas *et al.* (2006b), Rudelson and Vershynin (2007))**.**

$$\|\mathrm{A} - \tilde{\mathrm{A}}_k\|^2 \leq \|\mathrm{A} - \mathrm{A}_k\|^2 + 2\|\mathrm{A}^{\mathrm{T}}\mathrm{A} - \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\| \leq (\|\mathrm{A} - \mathrm{A}_k\| + \sqrt{2}\|\mathrm{A}^{\mathrm{T}}\mathrm{A} - \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\|^{1/2})^2.$$

Proof. The proof follows using standard arguments and an application of a perturbation theory result due to Weyl for bounding the change in any singular value upon hermitian perturbation of a hermitian matrix. ∎

Therefore, if we can approximate the matrix product $\mathrm{A}^{\mathrm{T}}\mathrm{A}$, we immediately get a good reconstruction for every k. The appropriate sampling probabilities from the previous section are

$$p_t \geq \beta \frac{\mathbf{a}_t^{\mathrm{T}}\mathbf{a}_t}{\|\mathrm{A}\|_F^2}.$$

In this case, if $r \geq (4\rho/\beta\epsilon^2)\ln\frac{2d}{\delta}$, then with probability at least $1-\delta$,

$$\|\mathrm{A} - \tilde{\mathrm{A}}_k\|^2 \leq \|\mathrm{A} - \mathrm{A}_k\|^2 + 2\epsilon\|\mathrm{A}\|^2.$$

The sampling probabilities are easy to compute and sampling can be accomplished in one pass if the matrix is stored row-by-row.

To get a relative error result, we need a more carefully constructed set of non-uniform sampling probabilities. The problem here becomes apparent if A has rank k. In this case we have no hope of a relative error approximation unless we preserve the rank during sampling. To do so, we need to sample according to the actual singular vectors in U, not according to A; this is because sampling according to A can give especially large weight to a few of the large singular value directions, ignoring the small singular value directions and hence not preserving rank. By sampling according to U, we essentially put equal weight on all singular directions. To approximate U well, we need sampling probabilities

$$p_t \geq \frac{\beta}{d}\mathbf{u}_t^{\mathrm{T}}\mathbf{u}_t.$$

Then, from Corollary 13, if $r \geq (4(d-\beta)/\beta\epsilon^2)\ln\frac{2d}{\delta}$, with probability at least $1-\delta$,

$$\|\mathrm{I} - \mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\| \leq \epsilon.$$

Since $\|\mathrm{U}\| = 1$, it also follows that

$$\|\mathrm{U}\mathrm{U}^{\mathrm{T}} - \mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\mathrm{U}^{\mathrm{T}}\| \leq \epsilon.$$

This result is useful because of the following lemma.

Lemma 20 (Spielman and Srivastava (2008))**.** *If* $\|\mathrm{U}\mathrm{U}^{\mathrm{T}} - \mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\mathrm{U}^{\mathrm{T}}\| \leq \epsilon$*, then for every* $\mathbf{x} \in \mathbb{R}^d$*,*

$$(1-\epsilon)\mathbf{x}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x} \leq \mathbf{x}^{\mathrm{T}}\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\mathbf{x} \leq (1+\epsilon)\mathbf{x}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x}.$$

Proof. We give a sketch of the proof from Spielman and Srivastava (2008). We let $\mathbf{x} \neq \mathbf{0}$ range over col(U). Since col(U) = col(A), $\mathbf{x} \in$ col(U) if and only if for some $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{x} = \mathrm{A}\mathbf{y}$. Since $\mathrm{rank}(\mathrm{A}) = d$, $\mathrm{A}\mathbf{y} \neq 0 \iff \mathbf{y} \neq 0$. Also note that $\mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{A} = \mathrm{A}$, since $\mathrm{U}\mathrm{U}^{\mathrm{T}}$ is a projection operator onto the column space of U, which is the same as the column space of A. The following sequence establishes the lemma.

$$\begin{aligned}
\|\mathrm{U}\mathrm{U}^{\mathrm{T}} - \mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\mathrm{U}^{\mathrm{T}}\| &= \sup_{\mathbf{x}\neq\mathbf{0}} \frac{|\mathbf{x}^{\mathrm{T}}\mathrm{U}\mathrm{U}^{\mathrm{T}}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\mathrm{U}^{\mathrm{T}}\mathbf{x}|}{\mathbf{x}^{\mathrm{T}}\mathbf{x}}, \\
&= \sup_{\mathrm{A}\mathbf{y}\neq\mathbf{0}} \frac{|\mathbf{y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{A}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{A}\mathbf{y}|}{\mathbf{y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{y}}, \\
&= \sup_{\mathrm{A}\mathbf{y}\neq\mathbf{0}} \frac{|\mathbf{y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{A}\mathbf{y}|}{\mathbf{y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{y}}, \\
&= \sup_{\mathbf{y}\neq\mathbf{0}} \frac{|\mathbf{y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\mathbf{y}|}{\mathbf{y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{y}},
\end{aligned}$$

The lemma now follows because $\|\mathrm{U}\mathrm{U}^{\mathrm{T}} - \mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\mathrm{U}\mathrm{U}^{\mathrm{T}}\| \leq \epsilon$. ■

Via the Courant-Fischer characterization Golub and Van Loan (1983) of the singular values, it is immediate from Lemma 20 that the singular value spectrum is also preserved :

$$(1-\epsilon)\sigma_i(\mathrm{A}^{\mathrm{T}}\mathrm{A}) \leq \sigma_i(\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}) \leq (1+\epsilon)\sigma_i(\mathrm{A}^{\mathrm{T}}\mathrm{A}). \tag{5}$$

Lemma 20 along with (5) will allow us to prove the relative approximation result.

Theorem 21. *If* $p_t \geq \frac{\beta}{d}\mathbf{u}_t^{\mathrm{T}}\mathbf{u}_t$ *and* $r \geq (4(d-\beta)/\beta\epsilon^2)\ln\frac{2d}{\epsilon}$*, then, for* $k = 1, \ldots, d$*,*

$$\|\mathrm{A} - \mathrm{A}\tilde{\Pi}_k\| \leq \left(\frac{1+\epsilon}{1-\epsilon}\right)^{1/2} \|\mathrm{A} - \mathrm{A}_k\|,$$

where $\tilde{\Pi}_k$ *projects onto the top* k *right singular vectors of* $\tilde{\mathrm{A}}$*.*

Remarks For $\epsilon \leq \frac{1}{2}$, $\left(\frac{1+\epsilon}{1-\epsilon}\right)^{1/2} \leq 1 + 2\epsilon$. Computing the probabilities p_t involves knowing $\mathbf{u}_t$ which means one has to perform an SVD, in which case, one could use A_k; it seems like overkill to compute A_k in order to approximate A_k. We discuss approximate sampling schemes later, in Section 7.

Proof. Let $\|\mathbf{x}\| = 1$. The following sequence establishes the result.

$$\begin{aligned}
\|\mathrm{A}(\mathrm{I} - \tilde{\Pi}_k)\|^2 &= \sup_{\mathbf{x}\in\ker(\tilde{\Pi}_k)} \|\mathrm{A}\mathbf{x}\|^2 = \sup_{\mathbf{x}\in\ker(\tilde{\Pi}_k)} \mathbf{x}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x}, \\
&\leq \frac{1}{1-\epsilon} \sup_{\mathbf{x}\in\ker(\tilde{\Pi}_k)} \mathbf{x}^{\mathrm{T}}\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\mathbf{x}, \\
&= \frac{1}{1-\epsilon}\sigma_{k+1}(\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}), \\
&\leq \frac{1+\epsilon}{1-\epsilon}\sigma_{k+1}(\mathrm{A}^{\mathrm{T}}\mathrm{A}) = \frac{1+\epsilon}{1-\epsilon}\|\mathrm{A} - \mathrm{A}_k\|^2.
\end{aligned}$$

■

6 ℓ_2 Linear Regression with Relative Error Bounds

A linear regression is represented by a real data matrix $\mathrm{A} \in \mathbb{R}^{m\times d}$ which represents m points in $\mathbb{R}^d$, and a target vector $\mathbf{y} \in \mathbb{R}^m$. Traditionally, $m \gg d$ (severly over constrained regression). The goal is to find a regression vector $\mathbf{x}^* \in \mathbb{R}^2$ which minimizes the ℓ_2 fit error (least squares regression)

$$\mathcal{E}(\mathbf{x}) = \|\mathrm{A}\mathbf{x} - \mathbf{y}\|_2^2 = \sum_{t=1}^{m}(\mathbf{a}_t^{\mathrm{T}}\mathbf{x} - y_t)^2,$$

We assume such an optimal $\mathbf{x}^*$ exists (it may not be unique unless A has full column rank), and is given by $\mathbf{x}^* = \mathrm{A}^+\mathbf{y}$, where $^+$ denotes the More-Penrose pseudo-inverse; this problem can be solved in $O(md^2)$. Through row-sampling, it is possible to construct $\hat{\mathbf{x}}$, an approximation to the optimal regression weights $\mathbf{x}^*$, which is a relative error approximation to optimal,

$$\mathcal{E}(\hat{\mathbf{x}}) \leq (1+\epsilon)\mathcal{E}(\mathbf{x}^*).$$

As usual, let $\mathrm{A} = \mathrm{U_A S_A V_A^T}$. Then $\mathrm{A}^+ = \mathrm{V_A S_A^{-1} U_A^T}$, and so $\mathbf{x}^* = \mathrm{V S^{-1} U^T y}$. The predictions are $\mathbf{y}^* = \mathrm{A}\mathbf{x}^* = \mathrm{U_A U_A^T y}$, which is the projection of $\mathbf{y}$ onto the column space of A. We define the residual $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{y}^* = \mathbf{y} - \mathrm{A}\mathbf{x}^* = (\mathrm{I} - \mathrm{U_A U_A^T})\mathbf{y}$, so

$$\mathbf{y} = \mathrm{U_A U_A^T}\mathbf{y} + \boldsymbol{\epsilon}. \tag{6}$$

We will construct $\tilde{\mathrm{A}}$ and $\tilde{\mathbf{y}}$ by sampling rows:

$$[\tilde{\mathrm{A}}, \tilde{\mathbf{y}}] = \mathrm{Q}[\mathrm{A}, \mathbf{y}],$$

and solve the linear regression problem on $(\tilde{\mathrm{A}}, \tilde{\mathbf{y}})$ to obtain $\hat{\mathbf{x}} = \tilde{\mathrm{A}}^+ \tilde{\mathbf{y}}$. For $\beta \in (0, \frac{1}{3}]$, we will use the sampling probabilities

$$p_t \geq \beta \left(\frac{\mathbf{u}_t^2}{d} + \frac{(\mathbf{u}_t^2 + \frac{\epsilon_t^2}{\boldsymbol{\epsilon}^\mathrm{T}\boldsymbol{\epsilon}})}{d+1} + \frac{\epsilon_t^2}{\boldsymbol{\epsilon}^\mathrm{T}\boldsymbol{\epsilon}} \right) \tag{7}$$

to construct $\tilde{\mathrm{A}}$ and $\tilde{\mathbf{y}}$. There are three parts to these sampling probabilities. The first part allows us to reconstruct A well from $\tilde{\mathrm{A}}$; the second allows us to reconstruct $\mathrm{A}^\mathrm{T}\boldsymbol{\epsilon}$; and, the third allows us to reconstruct $\boldsymbol{\epsilon}$.

Note that $\tilde{\mathrm{A}} = \mathrm{Q U_A S_A V_A}^\mathrm{T}$; if $\mathrm{QU_A}$ consisted of orthonormal columns, then this would be the SVD of $\tilde{\mathrm{A}}$. Indeed, this is approximately so, as we will soon see. Let the SVD of $\tilde{\mathrm{A}}$ be $\tilde{\mathrm{A}} = \mathrm{U_{\tilde{A}} S_{\tilde{A}} V_{\tilde{A}}}^\mathrm{T}$. Let $\tilde{\mathrm{U}} = \mathrm{QU_A}$. Since $p_t \geq \beta \mathbf{u}_t^2/d$, it follows from Corollary 13 that if $r \geq 2\frac{d-\beta}{\beta\epsilon^2}$, for $\epsilon \in (0,1)$, then, with high probability,

$$\|\mathrm{I} - \tilde{\mathrm{U}}^\mathrm{T}\tilde{\mathrm{U}}\| \leq \epsilon.$$

Since the eigenvalues of $\mathrm{I} - \tilde{\mathrm{U}}^\mathrm{T}\tilde{\mathrm{U}}$ are given by $1 - \sigma_i^2(\tilde{\mathrm{U}})$, it follows that

$$1 - \epsilon < \sigma_i^2(\tilde{\mathrm{U}}) < 1 + \epsilon.$$

So all the singular values of $\mathrm{U_A}$ are preserved after sampling. Essentially, it suffices to sample $r = O(d \ln d/\epsilon^2)$ rows to preserve the entire spectrum of $\mathrm{U_A}$. By choosing (say) $\epsilon = \frac{1}{2}$, the rank of $\mathrm{U_A}$ is preserved with high probability, since all the singular values are bigger than $\frac{1}{2}$. Thus, with high probability, $\mathrm{rank}(\tilde{\mathrm{A}}) = \mathrm{rank}(\mathrm{U_{\tilde{A}}}) = \mathrm{rank}(\mathrm{QU_A}) = \mathrm{rank}(\mathrm{U_A}) = \mathrm{rank}(\mathrm{A})$. Since $\mathrm{QU_A}$ has full rank, $\mathrm{S}_{\mathrm{QU_A}}^{-1}$ is defined, and $\mathrm{S_{QU_A}} - \mathrm{S}_{\mathrm{QU_A}}^{-1}$ is a diagonal matrix whose diagonals are $(\sigma_i^2(\tilde{\mathrm{U}}) - 1)/\sigma_i(\tilde{\mathrm{U}})$; thus, $\|\mathrm{S_{QU_A}} - \mathrm{S}_{\mathrm{QU_A}}^{-1}\|_2 \leq \epsilon/\sqrt{1-\epsilon}$. This allows us to quantify the degree to which $\mathrm{QU_A}$ is orthonormal, because

$$\begin{aligned} \|(\mathrm{QU_A})^+ - (\mathrm{QU_A})^\mathrm{T}\|_2 &= \|\mathrm{V_{QU_A} S_{QU_A}^{-1} U_{QU_A}}^\mathrm{T} - \mathrm{V_{QU_A} S_{QU_A} U_{QU_A}^T}\|_2 \\ &= \|\mathrm{S}_{\mathrm{QU_A}}^{-1} - \mathrm{S_{QU_A}}\|_2 \leq \frac{\epsilon}{\sqrt{1-\epsilon}}. \end{aligned}$$

Finally, we can get a convenient form for $\tilde{\mathrm{A}}^+ = (\mathrm{QA})^+$, because $\mathrm{QA} = \mathrm{QU_A S_A V_A^T}$ has full rank, and so $\mathrm{QU_A} = \mathrm{U_{QU_A} S_{QU_A} V_{QU_A}^T}$ has full rank (and hence is the product of full rank matrices). Thus,

$$\begin{aligned} (\mathrm{QA})^+ &= (\mathrm{U_{QU_A} S_{QU_A} V_{QU_A}^T S_A V_A^T})^+, \\ &= \mathrm{V_A}(\mathrm{S_{QU_A} V_{QU_A}^T S_A})^+ \mathrm{U_{QU_A}^T}, \\ &= \mathrm{V_A S_A^{-1} V_{QU_A} S_{QU_A}^{-1} U_{QU_A}^T}, \\ &= \mathrm{V_A S_A^{-1}}(\mathrm{QU_A})^+, \end{aligned}$$

We summarize all this information in the next lemma.

Lemma 22. *If $r \geq (4d/\beta\epsilon^2)\ln\frac{2d}{\delta}$, with probability at least $1-\delta$, all of the following hold:*

$$\operatorname{rank}(\tilde{\mathrm{A}}) = \operatorname{rank}(\mathrm{U}_{\tilde{\mathrm{A}}}) = \operatorname{rank}(\mathrm{QU_A}) = \operatorname{rank}(\mathrm{U_A}) = \operatorname{rank}(\mathrm{A}); \tag{8}$$

$$\|\mathrm{S}_{\mathrm{QU_A}} - \mathrm{S}^{-1}_{\mathrm{QU_A}}\|_2 \leq \epsilon/\sqrt{1-\epsilon}; \tag{9}$$

$$\|(\mathrm{QU_A})^+ - (\mathrm{QU_A})^{\mathrm{T}}\|_2 \leq \epsilon/\sqrt{1-\epsilon}; \tag{10}$$

$$(\mathrm{QA})^+ = \mathrm{V_A}\mathrm{S}_{\mathrm{A}}^{-1}(\mathrm{QU_A})^+. \tag{11}$$

In Lemma 22 we have simplified the constant to 4; this is a strengthened form of Lemma 4.1 in Drineas *et al.* (2006d); in particular, the dependence on d is near-linear.

Remember that $\hat{\mathbf{x}} = \tilde{\mathrm{A}}^+\tilde{\mathbf{y}}$; we now bound $\|A\hat{\mathbf{x}} - \mathbf{y}\|^2$. We only sketch the derivation which basically follows the line of reasoning in Drineas *et al.* (2006d). Under the conditions of Lemma 22, with probability at least $1-\delta$,

$$\begin{aligned}
\|\mathrm{A}\hat{\mathbf{x}} - \mathbf{y}\| &= \|\mathrm{A}\tilde{\mathrm{A}}^+\tilde{\mathbf{y}} - \mathbf{y}\| = \|\mathrm{A}(\mathrm{QA})^+\mathrm{Q}\mathbf{y} - \mathbf{y}\| \\
&\overset{(a)}{=} \|\mathrm{U_A}(\mathrm{QU_A})^+\mathrm{Q}\mathbf{y} - \mathbf{y}\| \\
&\overset{(b)}{=} \|\mathrm{U_A}(\mathrm{QU_A})^+\mathrm{Q}(\mathrm{U_A}\mathrm{U_A}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\epsilon}) - \mathrm{U_A}\mathrm{U_A}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\epsilon}\| \\
&\overset{(c)}{=} \|\mathrm{U_A}(\mathrm{QU_A})^+\mathrm{Q}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}\| \\
&= \|\mathrm{U_A}((\mathrm{QU_A})^+ - (\mathrm{QU_A})^{\mathrm{T}})\mathrm{Q}\boldsymbol{\epsilon} + \mathrm{U_A}(\mathrm{QU_A})^{\mathrm{T}}\mathrm{Q}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}\| \\
&\overset{(d)}{\leq} \|(\mathrm{QU_A})^+ - (\mathrm{QU_A})^{\mathrm{T}}\|\|\mathrm{Q}\boldsymbol{\epsilon}\| + \|\mathrm{U_A}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}\| \\
&\overset{(e)}{\leq} \frac{\epsilon}{\sqrt{1-\epsilon}}\|\mathrm{Q}\boldsymbol{\epsilon}\| + \|\mathrm{U_A}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}\|.
\end{aligned}$$

(a) follows from Lemma 22; (b) follows from (6); (c) follows Lemma 22, because $\mathrm{QU_A}$ has full rank and so $(\mathrm{QU_A})^+\mathrm{QU_A} = \mathrm{I}_d$; (d) follows from the triangle inequality and sub-multiplicativity using $\|\mathrm{U_A}\| = 1$; finally, (e) follows from Lemma 22. We now see the rationale for the complicated sampling probabilities. Since $p_t \geq \epsilon_t^2/\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}$, for r large enough, by Theorem 17, $\|\mathrm{Q}\boldsymbol{\epsilon}\|^2 \leq \|\boldsymbol{\epsilon}\|^2(1+\epsilon)$. Similarly, since $\mathrm{U_A}^{\mathrm{T}}\boldsymbol{\epsilon} = 0$, $\|\mathrm{U_A}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\boldsymbol{\epsilon}\| = \|\mathrm{U_A}^{\mathrm{T}}\boldsymbol{\epsilon} - \mathrm{U_A}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\boldsymbol{\epsilon}\|$; so, we can apply Lemma 14 with $\mathrm{S}_1 = \mathrm{I}_d$, $\mathrm{V} = \boldsymbol{\epsilon}/\|\boldsymbol{\epsilon}\|$ and $\mathrm{S}_2 = \|\boldsymbol{\epsilon}\|$. According to Lemma 14, if $p_t \geq \beta(\mathbf{u}_t^2 + \epsilon_t^2/\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon})/(d+1)$, then if r is large enough, $\|\mathrm{U_A}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\boldsymbol{\epsilon}\| \leq \epsilon\|\boldsymbol{\epsilon}\|$. Since these are all probabilistic statements, we need to apply the union bound to ensure that all of them hold. Ultimately, we have the claimed result:

Theorem 23. *For sampling probabilities satisfying (7), and for $r \geq (8(d+1)/\beta\epsilon^2)\ln\frac{2(d+1)}{\delta}$, let $\hat{\mathbf{x}} = (\mathrm{QA})^+\mathrm{Q}\mathbf{y}$ be the approximate regression. Then, with probability at least $1-3\delta$,*

$$\|\mathrm{A}\hat{\mathbf{x}} - \mathbf{y}\| \leq \left(1 + \epsilon + \epsilon\sqrt{\frac{1+\epsilon}{1-\epsilon}}\right)\|\mathrm{A}\mathbf{x}^* - \mathbf{y}\|,$$

where $\mathbf{x}^* = \mathrm{A}^+\mathbf{y}$ is the optimal regression.

Remarks For the proof of the theorem, we observe that any transformation matrix Q satisfying the following three properties with high probability will do:

$$(i)\|\mathrm{I} - \mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{QU}\| \leq \epsilon; \qquad (ii)\|\mathrm{Q}\boldsymbol{\epsilon}\| \leq (1+\epsilon)\|\boldsymbol{\epsilon}\|; \qquad (iii)\|\mathrm{U}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Q}\boldsymbol{\epsilon}\| \leq \epsilon\|\boldsymbol{\epsilon}\|.$$

7 Estimating the Spectral Norm

The row-norm based sampling is relatively straightforward for the symmetric product. For the asymmetric product, $\mathrm{A}^{\mathrm{T}}\mathrm{B}$, we need probabilities

$$p_t \geq \beta \frac{\frac{1}{\|\mathrm{A}\|^2}\mathbf{a}_t^{\mathrm{T}}\mathbf{a}_t + \frac{1}{\|\mathrm{B}\|^2}\mathbf{b}_t^{\mathrm{T}}\mathbf{b}_t}{\rho_{\mathrm{A}} + \rho_{\mathrm{B}}}. \tag{12}$$

To get these probabilities, we need $\|\mathrm{A}\|$ and $\|\mathrm{B}\|$; since we can compute the exact product in $O(md_1d_2)$, a practically useful algorithm would need to estimate $\|\mathrm{A}\|$ and $\|\mathrm{B}\|$ efficiently. Suppose we had estimates $\lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}$ which satisfy:

$$(1-\epsilon)\|\mathrm{A}\|^2 \leq \lambda_{\mathrm{A}}^2 \leq (1+\epsilon)\|\mathrm{A}\|^2; \qquad (1-\epsilon)\|\mathrm{B}\|^2 \leq \lambda_{\mathrm{B}}^2 \leq (1+\epsilon)\|\mathrm{B}\|^2.$$

We can construct probabilities satisfying the desired property with $\beta = (1-\epsilon)/(1+\epsilon)$.

$$\begin{aligned} p_t &= \frac{\frac{1}{\lambda_{\mathrm{A}}^2}\mathbf{a}_t^{\mathrm{T}}\mathbf{a}_t + \frac{1}{\lambda_{\mathrm{B}}^2}\mathbf{b}_t^{\mathrm{T}}\mathbf{b}_t}{\|\mathrm{A}\|_F^2/\lambda_{\mathrm{A}}^2 + \|\mathrm{B}\|_F^2/\lambda_{\mathrm{B}}^2} \\ &\geq \frac{\frac{1}{(1+\epsilon)\|\mathrm{A}\|^2}\mathbf{a}_t^{\mathrm{T}}\mathbf{a}_t + \frac{1}{(1+\epsilon)\|\mathrm{A}\|^2}\mathbf{b}_t^{\mathrm{T}}\mathbf{b}_t}{\|\mathrm{A}\|_F^2/(1-\epsilon)\|\mathrm{A}\|^2 + \|\mathrm{B}\|_F^2/(1-\epsilon)\|\mathrm{A}\|^2} \\ &= \left(\frac{1-\epsilon}{1+\epsilon}\right)\frac{\frac{1}{\|\mathrm{A}\|^2}\mathbf{a}_t^{\mathrm{T}}\mathbf{a}_t + \frac{1}{\|\mathrm{B}\|^2}\mathbf{b}_t^{\mathrm{T}}\mathbf{b}_t}{\rho_{\mathrm{A}} + \rho_{\mathrm{B}}}. \end{aligned}$$

One practical way to obtain $\|\mathrm{A}\|^2$ is using the power iteration. Given an arbitrary unit vector $\mathbf{x}_0$, for $n \geq 1$, let $\mathbf{x}_n = \mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x}_{n-1}/\|\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x}_{n-1}\|$. Note that multiplying by $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ can be done in $O(2md_1)$ operations. Since $\mathbf{x}_n$ is a unit vector, $\|\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x}_n\| \leq \|\mathrm{A}\|^2$. We now get a lower bound. Let $\mathbf{x}_0$ be a random isotropic vector constructed using d_1 independent standard Normal variates $z_1, \ldots, z_{d_1}$; so $\mathbf{x}_0^{\mathrm{T}} = [z_1, \ldots, z_{d_1}]/\sqrt{z_1^2 + \cdots + z_{d_1}^2}$. Let $\lambda_n^2 = \|\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x}_n\|$ be an estimate for $\|\mathrm{A}\|^2$ after n power iterations.

Lemma 24. *For some constant $c \leq (\frac{2}{\pi}+2)^3$, with probability at least $1-\delta$,*

$$\lambda_n^2 \geq \frac{\|\mathrm{A}\|^2}{\sqrt{4 + \frac{cd_1}{\delta^3}\cdot 2^{-2n}}}.$$

Remarks $n \geq c\log\frac{d_1}{\delta}$ gives the desired constant factor approximation. Since each power iteration takes $O(md_1)$ time, and we run $O(\log\frac{d_1}{\delta})$ power iterations, in $O(md_1\log\frac{d_1}{\delta})$ time, we obtain a sufficiently good estimate for $\|\mathrm{A}\|$ (and similarly for $\|\mathrm{B}\|$).

Proof. Assume that $\mathbf{x}_0 = \sum_{i=1}^{d_1}\alpha_i\mathbf{v}_i$, where $\mathbf{v}_i$ are the eigenvectors of $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ with corresponding eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_{d_1}^2$. Note $\|\mathrm{A}\|^2 = \sigma_1^2$. If $\sigma_{d_1}^2 \geq \sigma_1^2/2$, then it trivially follows that $\|\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x}_n\| \geq \sigma_1^2/2$ for any n, so assume that $\sigma_{d_1}^2 < \sigma_1^2/2$. We can thus partition the singular values into those at least $\sigma_1^2/2$ and those which are smaller; the latter set is non-empty. So assume for some $k < d_1$, $\sigma_k^2 \geq \sigma_1^2/2$ and $\sigma_{k+1}^2 < \sigma_1^2/2$. Since $\mathbf{x}_n = \sum_i \alpha_i\sigma_i^{2n}\mathbf{v}_i/(\sum_i \alpha_i^2\sigma_i^{4n})^{1/2}$, we therefore

have:

$$\begin{aligned}
\lambda_n^4 &= \|\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x}_n\|^2 = \frac{\sum_{i=1}^{d_1}\alpha_i^2\sigma_i^{4(n+1)}}{\sum_{i=1}^{d_1}\alpha_i^2\sigma_i^{4n}} \\
&\geq \frac{\sum_{i=1}^{k}\alpha_i^2\sigma_i^{4(n+1)}}{\sum_{i=1}^{d_1}\alpha_i^2\sigma_i^{4n}} \\
&= \frac{\sum_{i=1}^{k}\alpha_i^2\sigma_i^{4(n+1)}}{\sum_{i=1}^{k}\alpha_i^2\sigma_i^{4n}+\sum_{i=k+1}^{d_1}\alpha_i^2\sigma_i^{4n}}, \\
&= \sigma_1^4\frac{\sum_{i=1}^{k}\alpha_i^2(\sigma_i/\sigma_1)^{4(n+1)}}{\sum_{i=1}^{k}\alpha_i^2(\sigma_i/\sigma_1)^{4n}+\sum_{i=k+1}^{d_1}\alpha_i^2(\sigma_i/\sigma_1)^{4n}}, \\
&\overset{(a)}{\geq} \sigma_1^4\frac{\sum_{i=1}^{k}\alpha_i^2(\sigma_i/\sigma_1)^{4(n+1)}}{4\sum_{i=1}^{k}\alpha_i^2(\sigma_i/\sigma_1)^{4(n+1)}+2^{-2n}}, \\
&= \frac{\sigma_1^4}{4+2^{-2n}/\sum_{i=1}^{k}\alpha_i^2(\sigma_i/\sigma_1)^{4(n+1)}}, \\
&\overset{(b)}{\geq} \frac{\sigma_1^4}{4+2^{-2n}/\alpha_1^2}.
\end{aligned}$$

(a) follows because for $i \geq k+1$, $\sigma_i^2 < \sigma_1^2/2$; for $i \leq k$, $\sigma_1^2/\sigma_i^2 \leq 4$; and $\sum_{i\geq k+1}\alpha_i^2 \leq \sum_{i\geq 1}\alpha_i^2 = 1$. (b) follows because $\sum_{i=1}^{k}\alpha_i^2(\sigma_i/\sigma_1)^{4(n+1)} \geq \alpha_1^2$. The theorem will now follow if we show that with probability at least $1 - c\delta^{1/3}$, $\alpha_1^2 \geq \delta/d$. It is clear that $\mathbb{E}[\alpha_1^2] = 1/d$ from isotropy. Without loss of generality, assume $\mathbf{v}_1$ is aligned with the z_1 axis. So $\alpha_1^2 = z_1^2/\sum_i z_i^2$ ($z_1, \ldots, z_d$ are independent standard normals). For $\delta < 1$, we estimate $\mathbb{P}[\alpha_1^2 \geq \delta/d]$ as follows:

$$\begin{aligned}
\mathbb{P}\left[\alpha_1^2 \geq \frac{\delta}{d}\right] &= \mathbb{P}\left[\frac{z_1^2}{\sum_i z_i^2} \geq \frac{\delta}{d}\right] = \mathbb{P}\left[z_1^2 \geq \frac{\delta}{d}\sum_i z_i^2\right] = \mathbb{P}\left[z_1^2 \geq \frac{\delta}{d-\delta}\sum_{i\geq 2} z_i^2\right] \\
&\geq \mathbb{P}\left[z_1^2 \geq \frac{\delta}{d-1}\sum_{i\geq 2} z_i^2\right] \\
&\overset{(a)}{=} \mathbb{P}\left[\chi_1^2 \geq \frac{\delta}{d-1}\chi_{d-1}^2\right], \\
&\overset{(b)}{\geq} \mathbb{P}\left[\chi_1^2 \geq \delta + \delta^{2/3}\right] \cdot \mathbb{P}\left[\frac{\delta}{d-1}\chi_{d-1}^2 \leq \delta + \delta^{2/3}\right].
\end{aligned}$$

In (a) we compute the probability that a χ_1^2 random variable exceeds a multiple of an independent χ_{d-1}^2 random variable, which follows from the definition of the χ^2 distribution as a sum of squares of independent standard normals. (b) follows from independence and because one particular realization of the event in (a) is when $\chi_1^2 \geq \delta + \delta^{2/3}$ and $\delta\chi_{d-1}^2/(d-1) \leq \delta + \delta^{2/3}$. Since $\mathbb{E}[\chi_{d-1}^2/(d-1)] = 1$, and $Var[\chi_{d-1}^2/(d-1)] = 2/(d-1)$, by Chebyshev's inequality,

$$\mathbb{P}\left[\frac{\delta}{d-1}\chi_{d-1}^2 \leq \delta + \delta^{2/3}\right] \geq 1 - \frac{2\delta^{1/3}}{d-1}.$$

From the definition of the χ_1^2 distribution, we can bound $\mathbb{P}[\chi_1^2 \leq \delta + \delta^{2/3}]$,

$$\mathbb{P}[\chi_1^2 \leq \delta + \delta^{2/3}] = \frac{1}{2^{1/2}\Gamma(1/2)}\int_0^{\delta+\delta^{2/3}} du\ u^{-1/2}e^{-u/2} \leq \sqrt{\frac{2}{\pi}}(\delta + \delta^{2/3})^{1/2},$$

and so

$$\mathbb{P}\left[\alpha_1^2 \geq \frac{\delta}{d}\right] \geq \left(1 - \sqrt{\frac{2}{\pi}}(\delta + \delta^{2/3})^{1/2}\right) \cdot \left(1 - \frac{2\delta^{1/3}}{d-1}\right) \geq 1 - \left(\frac{2}{\pi} + 2\right)\delta^{1/3}.$$

■

We now consider the sampling based approach to estimate the spectral norm. Pre-sample the rows of A using probabilities proportional to the row norms to construct $\tilde{\mathrm{A}}$. We know that if $r \geq (4\rho_A/\beta\epsilon^2)\ln\frac{2d_1}{\delta}$, then

$$\|\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}} - \mathrm{A}^{\mathrm{T}}\mathrm{A}\| \leq \epsilon\|\mathrm{A}\|^2.$$

It follows that we have a ϵ-approximation to the spectral norm from

$$\begin{aligned}\|\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\| &= \|\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}} - \mathrm{A}^{\mathrm{T}}\mathrm{A} + \mathrm{A}^{\mathrm{T}}\mathrm{A}\| \leq (1+\epsilon)\|\mathrm{A}\|^2;\\ \|\mathrm{A}^{\mathrm{T}}\mathrm{A}\| &= \|\mathrm{A}^{\mathrm{T}}\mathrm{A} - \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}} + \tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\| \leq \epsilon\|\mathrm{A}\|^2 + \|\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\|.\end{aligned}$$

Thus, $(1-\epsilon)\|\mathrm{A}\|^2 \leq \|\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\| \leq (1+\epsilon)\|\mathrm{A}\|^2$. Along this route, one must first sample r rows, and then approximate the spectral norm of the resulting $\tilde{\mathrm{A}}$. We may now combine with the power iteration on $\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}$ to get a constant factor approximation efficiently (or we may compute exactly in $O(rd_1^2)$). Specifically, set $\epsilon = \frac{1}{2}$, in which case, with high probability, $\frac{1}{2}\|\mathrm{A}\|^2 \leq \|\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}\| \leq \frac{3}{2}\|\mathrm{A}\|^2$. Now, choose the number of power iterations $n \geq n^*$, where $\frac{cd_1}{\delta^3} = 2^{n^*}$. In this case, after n power iterations, we have an estimate which is at least $\frac{1}{\sqrt{5}}\|\tilde{\mathrm{A}}\|^2$ from Lemma 24, which proves Theorem 25.

Theorem 25. *With $r \geq (4\rho_A/\epsilon^2)\ln\frac{2d_1}{\delta}$, the spectral norm estimate $\tilde{\sigma}_1^2$ obtained after $c\ln\frac{d_1}{\delta}$ power iterations on $\tilde{\mathrm{A}}^{\mathrm{T}}\tilde{\mathrm{A}}$ starting from an isotropic random vector satisfies*

$$\frac{1}{2\sqrt{5}}\|\mathrm{A}\|^2 \leq \tilde{\sigma}_1^2 \leq \frac{3}{2}\|\mathrm{A}\|^2.$$

Further, the estimate $\tilde{\sigma}_1^2$ can be computed in $O(md_1 + \rho_A d_1/\epsilon^2 \ln^2(\frac{d_1}{\delta}))$.

As mentioned at the begining of this section, constant factor approximations to the spectral norms of the relevant matrices is enough to obtain probabilities satisfying (12) for some constant β.

References

Achlioptas, D. and McSherry, F. (2005). On spectral learning of mixtures of distributions. In *Proc. 18th Conference on Learning Theory (COLT)*, pages 458–469.

Achlioptas, D. and McSherry, F. (2007). Fast computation of low-rank approximations. *Journal of ACM*, **54**(2), Article 10.

Achlioptas, D., Fiat, A., Karlin, A., and McSherry, F. (2001). Web search via hub synthesis. In *Proc. 42nd IEEE symposium on Foundations of Computer Science*, pages 500–509.

Ahlswede, R. and Winter, A. (2002). Strong converse for identification via quantum channels. *Information Theory, IEEE Transactions on*, **48**(3), 569 –579.

Ailon, N. and Chazelle, B. (2006). Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proc. 38th ACM Symposium on Theory of Computing*, pages 557–563.

Azar, Y., Fiat, A., Karlin, A., McSherry, F., and Saia, J. (2001). Spectral analysis of data. In *Proc.33rd ACM symposium on Theory of computing*, pages 619–626.

Bernstein, S. (1924). On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine*, **4**(5), 38–49.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Rev.*, **37**(4), 573–595.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, **23**(4), 493–507.

Deshpande, A. and Vempala, S. (2006). Adaptive sampling and fast low-rank matrix approximation. In *Proc. 10th RANDOM*.

Deshpande, A., Rademacher, L., Vempala, S., and Wang, G. (2006). Matrix approximation and projective clustering via volume sampling. In *Proc. 17th ACM-SIAM symposium on Discrete algorithms*, pages 1117–1126.

Drineas, P., Kerenidis, I., and Raghavan, P. (2002). Competitive recommendation systems. In *Proc. 34th ACM symposium on Theory of computing*, pages 82–90.

Drineas, P., Frieze, A., Kannan, R., Vempala, S., and Vinay, V. (2004). Clustering large graphs via the singular value decomposition. *Machine Learning*, **56**(1-3), 9–33.

Drineas, P., Kannan, R., and Mahoney, M. W. (2006a). Fast Monte Carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, **36**(1), 132–157.

Drineas, P., Kannan, R., and Mahoney, M. W. (2006b). Fast Monte Carlo algorithms for matrices ii: Computing a low rank approximation to a matrix. *SIAM Journal on Computing*, **36**(1), 158–183.

Drineas, P., Kannan, R., and Mahoney, M. W. (2006c). Fast Monte Carlo algorithms for matrices iii: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, **36**(1), 184–206.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. (2006d). Sampling algorithms for l2 regression and applications. In *Proc. 17th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1127–1136.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. (2006e). Subspace sampling and relative-error matrix approximation: Column-row based methods. In *Proc. 14th ESA*.

Drineas, P., Magdon-Ismail, M., and Mahoney, M. (2010). Sampling leverage scores.

Frieze, A., Kannan, R., and Vempala, S. (1998). Fast monte-carlo algorithms for finding low-rank approximations. In *Proc. 39th Annual Symposium on Foundations of Computer Science*, pages 370–378.

Golub, G. and Van Loan, C. (1983). *Matrix Computations*. Johns Hopkins University Press, Baltimore.

Golub, G. and van Loan, C. (1996). *Matrix computations*. The Johns Hopkins University Press, London, 3 edition.

Gross, D., Liu, Y.-K., Steven Flammia, S. T., Becker, S., and Eisert, J. (2009). Quantum state tomography via compressed sensing. preprint available at: arXiv:0909.3304v2.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**(301), 13–30.

Kannan, R., Salmasian, H., and Vempala, S. (2008). The spectral method for general mixture models. *SIAM Journal of Computing*, **38**(3), 1141–1156.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**(5), 604–632.

Kuczyński, J. and Woźniakowski, H. (1989). Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. Technical Report CUCS-026-92, Institute of Informatics, University of Warsaw.

Magdon-Ismail, M. (2010). Row sampling for matrix algorithms via a non-commutative bernstein bound. *CoRR*, **abs/1008.0587**.

Magen, A. and Zouzias, A. (2010). Low rank matrix-valued chernoff bounds and applications. *submitted*. http://arxiv.org/abs/1005.2724.

McSherry, F. (2001). Spectral partitioning of random graphs. In *Proc 42nd IEEE symposium on Foundations of Computer Science*, pages 529–537.

Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, **61**(2), 217–235.

Recht, B. (2009). A simpler approach to matrix completion. working paper.

Rudelson, M. and Vershynin, R. (2007). Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, **54**(4), 19.

Sarlos, T. (2006). Improved approximation algorithms for large matrices via random projections. In *Proc. 47th IEEE Symposium on Foundations of Computer Science*, pages 143–152.

Spielman, D. A. and Srivastava, N. (2008). Graph sparsification by effective resistances. In *STOC '08: Proc.40th ACM Symposium on Theory of Computing*, pages 563–568.

Wigderson, A. and Xiao, D. (2008). Derandomizing the ahlswede-winter matrix-valued chernoff bound using pessimistic estimators, and applications. *Theory of Computing*, **4**(3), 53–76.

Woolfe, F., Liberty, E., Rokhlin, V., , and Tygert, M. (2008). A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, **25**(3), 335–366.